XTREME

A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization

Michal Hala
469265

Faculty of Informatics, Masaryk University

May 18, 2021

Contents

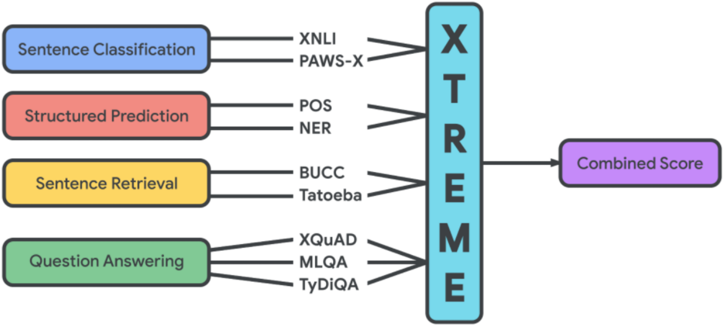
- What is XTREME?
- Design principles
- Tasks
- Languages
- Training and baselines
- Analysis of results

Motivation

- We have a language model gained by machine learning.
- How good is the model at solving different tasks?
- How well would the model work with different languages?

XTREME

- Cross-lingual TRansfer Evaluation of Multilingual Encoders.
- Benchmark for evaluating language models on 9 different tasks in 40 languages.



XTREME

- Most of the existing benchmarks work only with English for which the results are already comparable with humans.
- Cross-lingually transferred models are still relatively weak.
- Hardest problems are syntactic and sentence retrieval tasks.
- Solving tasks for some languages is more difficult than for others.

Design principles

- Tasks must be difficult enough.
- Tasks must be diverse.
- Training must not take too long (<1 day).
- The more languages the better.
- Languages should have enough monolingual data.
- Accessibility through public licences.

Tasks

- Sentence classification
  - XNLI (entailment and contradiction detection)
  - PAWS-X (paraphrase detection)
- Structured prediction
  - POS (part-of-speech tagging)
  - NER (name entity recognition)
- Sentence retrieval
  - BUCC (extraction of parallel sentences)
  - Tatoeba (similiarity of parallel sentences)
- Question answering
  - XQuAD
  - MLQA
  - TyDiQA-GoldP

Languages

- XTREME uses 40 languages from 12 language families.
- Most texts are retrieved from Wikipedia articles.
- Supported languages: af, ar, bg, bn, de, el, en, es, et, eu, fa, fi, fr, he, hi, hu, id, it, ja, jv, ka, kk, ko, ml, mr, ms, my, nl, pt, ru, sw, ta, te, th, tl, tr, ur, vi, yo, and zh.

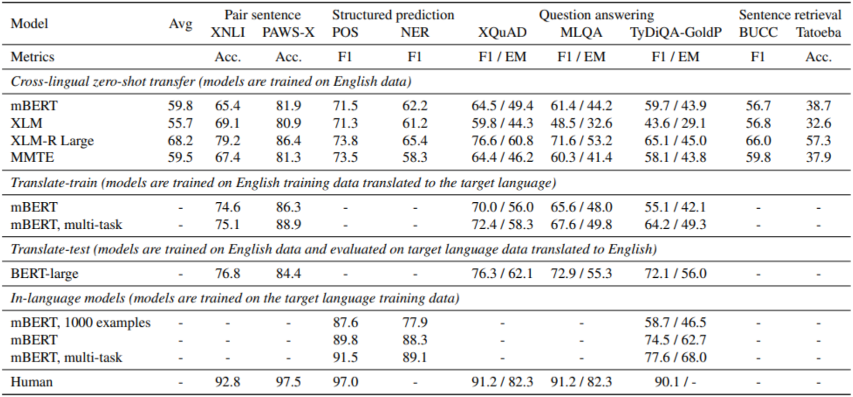
Training

- Uses zero-shot cross-lingual transfer with English as the core language.
- Model is fine tuned on English data.
- Applied on multilingual data afterwards.

Baselines

- XTREME presents several baseline models against which a user can compare their own model.
  - mBERT
  - XLM
  - XLM-R
  - MMTE

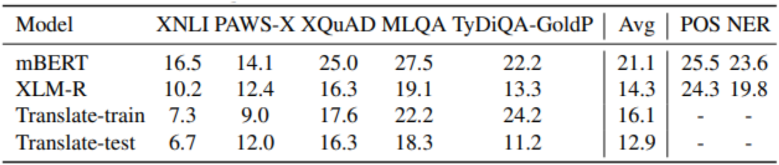
Results



Model	Avg	Pair sentence		Structured prediction		Question answering			Sentence retrieval	
		XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA-GoldP	BUCC	Tatoeba
Metrics		Acc.	Acc.	F1	F1	F1 / EM	F1 / EM	F1 / EM	F1	Acc.
Cross-lingual zero-shot transfer (models are trained on English data)										
mBERT	59.8	65.4	81.9	71.5	62.2	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9	56.7	38.7
XLM	55.7	69.1	80.9	71.3	61.2	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1	56.8	32.6
XLM-R Large	68.2	79.2	86.4	73.8	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	66.0	57.3
MMTE	59.5	67.4	81.3	73.5	58.3	64.4 / 46.2	60.3 / 41.4	58.1 / 43.8	59.8	37.9
Translate-train (models are trained on English training data translated to the target language)										
mBERT	-	74.6	86.3	-	-	70.0 / 56.0	65.6 / 48.0	55.1 / 42.1	-	-
mBERT, multi-task	-	75.1	88.9	-	-	72.4 / 58.3	67.6 / 49.8	64.2 / 49.3	-	-
Translate-test (models are trained on English data and evaluated on target language data translated to English)										
BERT-large	-	76.8	84.4	-	-	76.3 / 62.1	72.9 / 55.3	72.1 / 56.0	-	-
In-language models (models are trained on the target language training data)										
mBERT, 1000 examples	-	-	-	87.6	77.9	-	-	58.7 / 46.5	-	-
mBERT	-	-	-	89.8	88.3	-	-	74.5 / 62.7	-	-
mBERT, multi-task	-	-	-	91.5	89.1	-	-	77.6 / 68.0	-	-
Human	-	92.8	97.5	97.0	-	91.2 / 82.3	91.2 / 82.3	90.1 / -	-	-

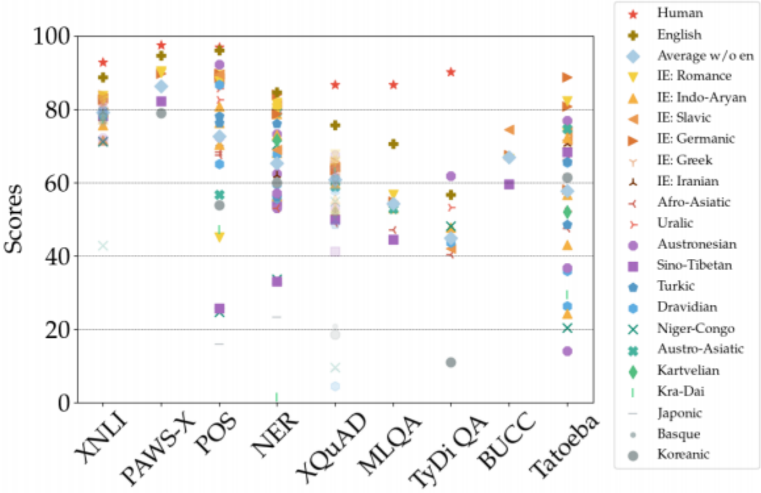
Cross-lingual transfer gap

- Measured difference between the performance on the English test set and all other languages (the lower the better).

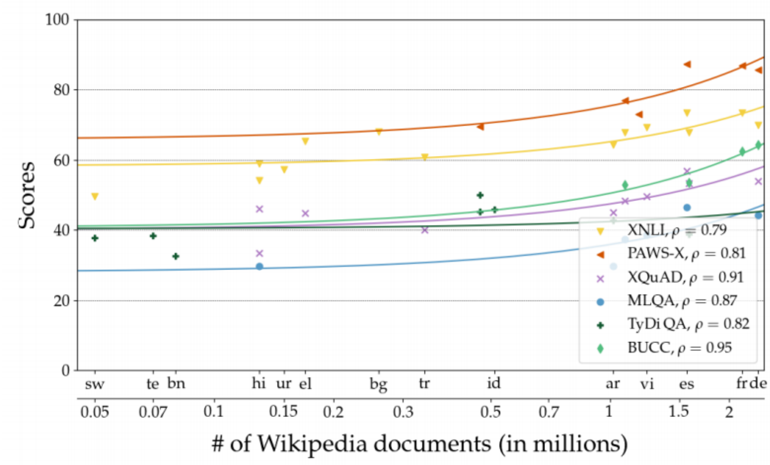


Model	XNLI	PAWS-X	XQuAD	MLQA	TyDiQA-GoldP	Avg	POS	NER
mBERT	16.5	14.1	25.0	27.5	22.2	21.1	25.5	23.6
XLM-R	10.2	12.4	16.3	19.1	13.3	14.3	24.3	19.8
Translate-train	7.3	9.0	17.6	22.2	24.2	16.1	-	-
Translate-test	6.7	12.0	16.3	18.3	11.2	12.9	-	-

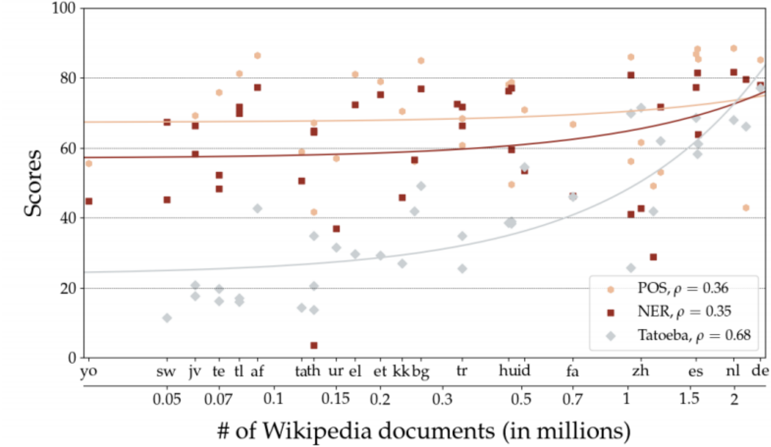
Language difference



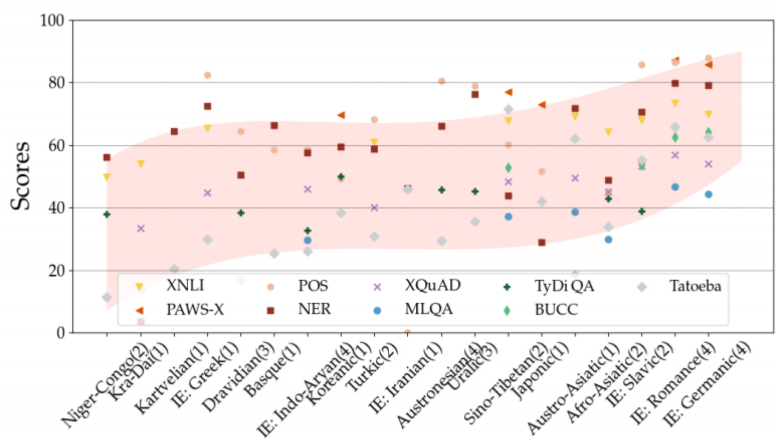
Correlation between training data size and score I



Correlation between training data size and score II



mBERT score on language families



Sources

- Hu, J. et al.: XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, url: https://arxiv.org/abs/2003.11080
- https://github.com/google-research/xtreme

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS